Rise of the Underdog

Heather Ford

Author's pre-print to appear in Wikipedia @20. Edited by Joseph Reagle and Jackie Koerner. Published by MIT Press (in press, expected early 2020) See https://reagle.org/joseph/2018/wp/at-20.html for more information about the book.

We all love an underdog. And when *Nature* announced that Wikipedia's quality was almost as good as *Encyclopaedia Britannica* for articles about science in 2005, I celebrated. I celebrated because Wikipedia was the David to Big Media's Goliath—the little guy, the people's encyclopaedia, the underdog who had succeeded against all odds.

Since then, Wikipedia has moved from thirty-seventh to fifth most visited website in the world. Wikipedia is now the top dog for facts as the world's most powerful platforms use to power their question and answer systems. Those platforms extract information from Wikipedia articles to fuel the question and answer systems that drive search engines like Google search and digital assistants including Amazon's Alexa, Apple's Siri and Google's Assistant.

It is tempting to see Wikipedia in 2020 as the new top dog in the world of facts. The problem is that Wikipedia's status is dependent almost entirely on Google, and the ways in which Wikipedia's content is increasingly represented without credit by major platforms signals Wikipedia's greatest existential threat to date.

Removing links back to Wikipedia as the source of answers to user questions prevents users from visiting Wikipedia in order to donate or volunteer. More important, however, are the ways in which unattributed facts violate the principle of verifiability on which Wikipedia was founded.

Within the bounds of Wikipedia, users are able to question whether statements are correctly attributed to reliable sources. They are able to contribute to discussions towards consensus, and to recognise the traces that signal how unstable or stable statements of fact are. But when those statements are represented without attribution or links back to their messy political and social contexts, they appear as the objective, natural and stable truth.

In 2020, there are new Goliaths in town in the form of the world's most powerful technology companies, and Wikipedia must re-articulate its foundational principles and highlight its underdog status if it wishes to reinstitute itself as a bastion of justice on the internet.

Once the Underdog

The underdog is a common archetype of some of the most enduring narratives—from the world of sport to politics. Studying the appeal of underdogs over a number of years, Vandello, Goldschmied, and Michniewicz define underdogs as "disadvantaged parties facing advantaged opponents and unlikely to succeed."[1] They write that there are underdog stories from cultures around the world: from the story of David and Goliath, in which the smaller David fights and kills the giant, Goliath, to the Monkey and the Turtle, a Philippine fable in which the patient turtle outwits the physically stronger and selfish monkey.

Underdogs are appealing because they offer an opportunity for redemption—a chance for the weaker individual or group to face up against a stronger opponent and to beat them, despite the odds leaning significantly against them. Usually, underdogs face off to better resourced competitors in a zero-sum game such as an election or sporting match, but underdogs don't need to win to be appealing. As Vandello et al state: they just have to *face up* to the bigger, more powerful, better resourced competitor in order to win the hearts of the public.

With the headline "Internet encyclopaedias go head to head" the Nature study represented such a competition when it was published in 2005.[2] The study pitted a four-year-old Wikipedia against the centuries-old Britannica by asking academic experts to compare 42 articles relating to science. The verdict? The average science entry in Wikipedia contained around four inaccuracies to Britannica's three, leading Nature to announce that "Jimmy Wales' Wikipedia comes close to Britannica in terms of the accuracy of its science entries."

The Nature study is now the stuff of legend. Although it was criticized for the way that articles were compared and the way that the study was reported, it is mostly used as evidence of the quality of Wikipedia in comparison to traditionally authored reference works.[3] For those of us working in the free and open source software and open content movement, it confirmed what we already thought we knew: that online resources like Wikipedia could attain the same (if not greater) level of quality that traditionally published resources enjoyed because they were open for the public to improve. It gave credence to the idea that content as well as software benefited from openness because, as Eric Raymond famously wrote, "with enough eyes, all bugs are shallow."[4]

In 2005, Wikipedia was being developed on the back of volunteer labour, a handful of paid employees and a tiny budget. In 2005, I was deep into my tenure as a digital commons activist. As the public lead for Creative Commons South Africa, the Executive Director of iCommons and the advisory board of the Wikimedia Foundation, I was in the business of selling openness to the world. In the photographs from 2005, I see myself smiling, surrounded by like-minded people from around the world who would meet at the annual iCommons Summit or Wikimania. We would talk about how copyleft was critical to a more innovative Internet. For me, freedom and openness via copyleft licenses provided the opportunity for greater access to educational materials critical for countries like my own, burdened by extreme copyright regimes that benefitted corporate publishing houses outside of South Africa at the expense of access to knowledge. I believed that open content and free and open source software was in keeping with the sharing of culture emblematic

of ubuntu, the Zulu and Xhosa term for "humanity towards others"—the belief in a universal bond that connects people around the world.

Life as an "Internet rights activist" wasn't all glamorous. Back home in Johannesburg, it meant countless meetings with anyone who would listen. Talking to funders, academics, lawyers, musicians, publishers, authors, we would present copyleft as an obvious choice for public knowledge, creativity, education and creative industries to tiny audiences of sceptical and/or curious individuals. In my case, it meant tears of frustration when debating intellectual property lawyers about the virtues of the South African Constitutional Court's finding in favour of the Trade Mark dispute between a young t-shirt producer and a multinational beer company. And righteous indignation when hearing about underhand attempts by large software corporations to stem the tide of open source in order to protect their hold on public education in Namibia.

I celebrated Wikipedia's success because it was a signal from the establishment that openness was a force to be recognised. I celebrated because Wikipedia had become emblematic of the people of the internet's struggle against Big Media. It signalled success against corporate media giants like the Motion Picture Association of America and its members who were railing aggressively against the ideology and practice of free and open source software and open content because it was considered a significant threat to their business models. In 2005, P2P firms, Napster, Grokster and StreamCast had been successfully sued by rights holders and Lawrence Lessig had lost his case to prevent US Congress from extending US copyright terms. We all needed a hero and we needed a few wins under our righteous belts.

When the Nature study was published in 2005, Wikipedia represented "the people of the Internet" against an old (and sizeable) Big Media who railed against any change that would see them threatened. Ironically, the company behind Encyclopedia Britannica was actually ailing when the Nature study drove the final nail into its coffin. But no matter: Britannica represented the old and Wikipedia the new. A year later, in 2006, Time Magazine's Person of the Year reinforced this win. Awarding the Person of the Year to "you", the editorial argued that ordinary people now controlled the means of producing information and media because they dissolved the power of the gatekeepers who had previously controlled the public's access to information.

> [2006 is] a story about community and collaboration on a scale never seen before. It's about the cosmic compendium of knowledge Wikipedia and the million-channel people' s network YouTube and the online metropolis MySpace. It's about the many wresting power from the few and helping one another for nothing and how that will not only change the world, but also change the way the world changes.[5]

It is this symbolic value that makes underdogs so powerful. Vandello, Goldschmied, and Michniewicz argue that we root for underdogs, not only because we want them to succeed but because we feel "it is right and just for them to do so." We dislike the fact that there is inequality in society—that some individuals or groups face a much more difficult task because they are under-resourced. Rooting for the underdog enables us to reconcile or face this injustice (albeit from a distance).

Wikipedia Wars

With few resources and Big Media set against them, Wikipedia was once seen as the underdog to traditional media. As the bastion of openness against the selfishness of proprietary media, its fight was seen as a just one. This was 15 years ago and now much is changed.

The encyclopedia that was pitted as Wikipedia's competitor, Britannica, is now all but dead (the final print version was published in 2010). Wikipedia has moved from 37th most visited website in the world when Nature published its study in 2005 to fifth place and enjoys about 18 billion pageviews a month.

Donations to Wikipedia's host non-profit, the Wikimedia Foundation's increased dramatically—from about $1.5 million in 2006 to almost $100 million in 2018. From a tiny office in a shopping mall in Atlanta with three employees to corporate headquarters in the heart of San Francisco and a staff of almost 260, the Wikimedia Foundation's operating budget and cash reserves are so healthy that some have argued that Wikipedia doesn't need your donations and that the increased budget is turning the Foundation into a corporate behemoth that is unaccountable to its volunteers.[6]

If there is a political battle being fought—between politicians, policies, ideologies or identities—there will be a parallel conflict on Wikipedia. On English Wikipedia, for example, Donald Trump's page is in a constant state of war. In 2018, an edit war ensued about whether to include information about Trump's performance at the 2018 US-Russia summit in Helsinki[7]. On the Brexit article, editors have received death threats and doxx attempts when editing information about the impact of Brexit on the UK and Europe[8]. After Time Magazine published a story by Aatish Tasser critical of Indian Prime Minister, Narendra Modi, Tasser's English Wikipedia page was vandalised and screenshots of the vandalised page distributed over social media as evidence[9].

The above examples relate to obviously political subjects, but Wikipedia wars are being fought beyond the bounds of politicians' biographies. Representation of current events on Wikipedia is almost always hotly contested. For almost every terrorist attack, natural disaster, or political protest, there will be attempts by competing groups to wrest control over the event narrative on Wikipedia in order to reflect their version of what happened, to whom it happened and why it happened. Unexpected events have consequences—for victims, perpetrators and the governments who distribute resources as a result of such classifications. Wikipedia is therefore regularly the site of battles over what becomes recognised as the neutral point of view, the objective fact, the common sense perspectives that affect the decisions that ultimately determine who the winners and losers are in the aftermath of an event.

Because of Wikipedia's growing authority, governments now block the site in order to prevent it from being used to distribute what they deem to be subversive ideas. Wikipedia is currently blocked in China and Turkey, but countries including France, Iran, Pakistan, Russia, Thailand, Tunisia, the United Kingdom and Venezuela have blocked specific content from a period of a few days to many years.

In 2013, it was found that Iran's censorship of Persian Wikipedia targeted a wide breadth of political, social, religious and sexual themes including information related to the Iranian government's human rights record and individuals who have challenged authorities.[10] In the UK, the Wikipedia article about "Virgin Killer", an album by the German rock band, Scorpions, was blacklisted for three days by the Internet Watch Foundation when the album cover image was classified as child pornography. In early 2019, all language editions of Wikipedia were blocked in Venezuela probably because of a Wikipedia article that listed newly-appointed National Assembly president Juan Guaidó as "president number 51 of the Bolivarian Republic of Venezuela", thus challenging Nicolás Maduro' s presidency.[11]

How has representation on Wikipedia come to matter so much? The answer is that Wikipedia matters more in the context of the even more powerful third party platforms that make use of its content, than the way it represents subjects on its articles. What matters most is not so much how facts are represented on Wikipedia but about how facts that originate on Wikipedia travel to other platforms.

Ask Google who the President of Uganda is who won MasterChef Australia last year and the results will probably be sourced from (English) Wikipedia in a special "knowledge panel" featured on the right hand side of the search results and in featured snippets at the top of organic search results. Ask Siri the same questions, and she will probably provide you with an answer that was originally extracted as data from Wikipedia.

Information in Wikipedia articles is being increasingly datafied and extracted by third parties in order to feed a new generation of question answer machines. If one can control how Wikipedia defines and represents a person, place, event or thing, then one can control how it is represented not only on Wikipedia but on Google, Apple, Amazon and other major platforms. This has not gone unnoticed by the many search engine optimizers, marketers, public relations and political agents who send their agents to do battle over facts on Wikipedia.

New Goliaths

From all appearances, then, Wikipedia is now the top dog in the world of facts. Look a little deeper into how Wikipedia arrived at this point and what role it is playing in the new Web ecosystem, however, and the picture becomes a little muddier. Britannica may be dead and Wikipedia may be the most popular encyclopedia, but Wikipedia is now more than just an encyclopedia and there are new Goliaths on which Wikipedia is so dependent for its success that could very easily wipe Wikipedia off the face of the internet.

Google has always prioritised Wikipedia entries in search results, and this is the primary way through which users have discovered Wikipedia content. But in 2012, Google announced a new project that would change how it organised search results. In a blog post entitled "Things not strings" VP of Engineering for Google, Amit Singhal wrote that Google was using Wikipedia and other public data sources to seed a Knowledge Graph that would provider "smarter search results" for users. [12] In addition to returning a list of possible results including Wikipedia articles when a user searched for "Marie Curie", for example, Google would present a "knowledge panel" on the right hand side of the page that would

"summarize relevant content around that topic, including key facts you're likely to need for that particular thing."

Soon after Google's announcement, former Head of Research at the Wikimedia Foundation, Dario Taraborelli started taking notice of how Google represented information from Wikipedia in its knowledge panels. One of the first iterations featured a prominent backlink to Wikipedia and even the Creative Commons Attribution Share-Alike license that Wikipedia content is licensed under. But, as the panels evolved, blue links to Wikipedia articles started shrinking in size. Over time, the underscore was removed so that the links weren't clickable, and then the links were lightened to a barely visible grey tone.

Taraborelli was concerned at how dependent Wikipedia was on Google and at how changes that were being made to the way that Wikipedia content was being presented by the search giant could have a significant impact on the sustainability of Wikipedia. If users were being presented with information from Wikipedia without having to visit the site, or without even knowing that Wikipedia was the true source, then that would surely affect the numbers of users visiting Wikipedia—as readers, editors or contributors to the annual fundraising campaign. These fears were confirmed by research conducted by McMahon et al who found that facts in the knowledge panels were being predominantly sourced from Wikipedia but that these were "almost never cited" and that this was leading to a significant reduction in traffic to Wikipedia.[13]

Taraborelli was also concerned with a more fundamental principle at issue here: that Google's use of Wikipedia information without credit "undermines people's ability to verify information and, ultimately, to develop well-informed opinions."[14] Verifiability is one of Wikipedia's core content policies. It is defined as the ability for "readers (to be) able to check whether information within Wikipedia articles is not just made up."

For editors, verifiability means that "all material must be attributable to reliable, published sources." Wikipedia's verifiability policy, in other words, establishes rights for readers and responsibilities for editors. Readers should have the right to be able to check whether information from Wikipedia is accurately represented by the reliable source from which it originates. Editors should ensure that all information should be attributable to reliable sources and that information that is likely to be challenged should be attributed using in-text citations.

It is easy to see Wikipedia as a victim of Google's folly here. The problem is that a project within the Wikimedia stable, Wikidata, has done exactly the same thing—as Andreas Kolbe pointed out in response to the Washington Post story about Google's knowledge boxes.[15] Launched to help efforts just like Google to better represent Wikipedia's facts by serving as a central storage of structured data for Wikimedia projects, Wikidata has been populated by millions of statements that are either uncredited to a reliable source or attributed to the entire Wikipedia language version from where they were extracted. The latter does not meet the requirements for verifiability, one of Wikipedia's foundational principles, because it does not enable downstream users the ability to verify or check whether the statements are, indeed, reflective of their source or whether the source itself is reliable or not.

A number of Wikipedians have voiced concern over Wikidata's apparent unconcern with the need for accurate source information for its millions of claims. Andreas Kolbe has contributed multiple articles about the problems with Wikidata. He wrote an op ed about Wikidata in December of 2015 as a counterpoint to the celebratory piece that had been published about the project the month prior.[16]

Kolbe made three observations about the quality of content on Wikidata. The first was the problem of unreferenced or under-referenced claims (more than half of the claims at that time were unreferenced). Second was the fact that Wikidata was extracting facts from Wikipedia and then presenting them under a more permissible copyright license than that of Wikipedia which was giving the green light to third party users like Google to use that content unattributed. And third, that there were problems with the quality of information on Wikidata because of its lack of stringent quality controls.

Kolbe noted a list of "Hoaxes long extinguished on Wikipedia live on, zombie-like, in Wikidata." Wikidata represents a strategic opportunity for search engine optimisation specialists and public relations professionals to influence search results. Without stringent quality control mechanisms, however, inaccurate information could be replicated and mirrored on more authoritative platforms which would multiply their detrimental effects.

In the past few years, the list of major platforms making use of Wikipedia information (either directly or via Wikidata) has grown. The most important re-users are now digital assistants in the form of Amazon's Alexa, Apple's Siri and Google's Assistant who answer user questions authoritatively using Wikipedia information. The loss of citations and links back to Wikipedia have grown alongside them, as problems of citation loss with Google and Wikidata have been replicated.

The problem, then, is about the process of automated extraction and the logics of knowledge bases more generally, than it is about the particular practices by specific companies or organisations. In 2015 and 2016, I wrote a series of articles about this problem with Mark Graham from the Oxford Internet Institute when I was a PhD student there. We argued that the process of automation in the context of the knowledge base had both practical and ethical implications for Internet users. [17]

From a practical perspective, we noted that information became less nuanced and its provenance or source obscured. The ethical case involved the loss of agency by users to contest information when that information is transported to third parties like Google. When incorrect information is not linked back to Wikipedia, users are only able to click on a link. There are no clear policies on how information can be changed or who is accountable for that information.

In one case, a journalist whose information was incorrectly appearing in the knowledge panel was informed by Google to submit feedback from multiple IP addresses, every 3 or 4 days multiple times, using different logins and to "get more people to help you submit feedback."[18] This does not constitute a policy on rectifying false information. Compare Wikipedia's editorial system with its transparent (albeit multitidudinous) policies and one realises how the datafication of Wikipedia content has removed important rights from internet users.

The Right to Verifiability

Wikipedia was once celebrated because it was seen as the underdog to Big Media. As Wikipedia has become increasingly powerful as a strategic resource for the production of knowledge about the world, battles over representation of its statements have intensified. Wikipedia is strategic today, not only because of how people, places, events and things are represented in its articles, but because of the ways in which those articles have become fodder for search engines and digital assistants. From its early prioritisation in search results, Wikipedia's facts are now increasingly extracted without credit by Artificial Intelligence processes that consume its knowledge and present it as objective fact.

As the fifth most popular website in the world, it is tempting in 2020 to see Wikipedia as a top dog in the world of facts but the consumption of Wikipedia's knowledge without credit introduces Wikipedia's greatest existential threat to date. This is not only because of the ways in which third party actors appropriate Wikipedia content and remove the links that might sustain the community in terms of contributions of donations and volunteer time. More important is that unsourced Wikipedia content threatens the principle of verifiability, one of the fundamental principles on which Wikipedia was built.

Verifiability sets up a series of rights and obligations by readers and editors of Wikipedia to knowledge whose political and social status is transparent. By removing direct links to the Wikipedia article where statements originate from, search engines and digital assistants are removing the clues that readers could use to a) evaluate the veracity of claims and b) take active steps to change that information through consensus if they feel that it is false. Without the source of factual statements being attributed to Wikipedia, users will see those facts as solid, incontrovertible truth, when, in reality, they may have been extracted in the midst of a process of consensus building or at the moment in which the article was vandalised.

Until now, platform companies have been asked to contribute to the Wikimedia Foundation's annual fundraising campaign in order to "give back" to what they are taking out of the commons.[19] But contributions of cash will not solve what amounts to Wikipedia's greatest existential threat to date. What is needed is a public campaign to reinstate the principle of verifiability in the content that is extracted from Wikipedia by platform companies. Users need to be able to understand a) exactly where facts originate b) how stable or unstable those statements are, c) how they might become involved in improving the quality of that information and d) the rules under which decisions about representation will be made.

Wikipedia was once recognised as the underdog because it was both under-resourced but, more importantly, because it represented the just fight against more powerful media who sought to limit the possibilities of people around the world to build knowledge products together. Today, the fight is a new one and Wikipedia must adapt in order to survive.

Sitting back and allowing platform companies to ingest Wikipedia's knowledges and represent it as the incontrovertible truth, rather than the messy and variable truths it actually stands in for, is an injustice. It is an injustice not only for Wikipedians but for

people around the world who use the resource – either directly on Wikimedia servers or indirectly via other platforms like search.

Notes

1. Joseph A. Vandello, Nadav Goldschmied, and Kenneth Michniewicz, "Underdogs as Heroes," in *Handbook of Heroism and Heroic Leadership*, ed. Allison, Scott T., George R. Goethals, and Roderick M. Kramer. (New York: Routledge, 2017), 339–355.

2. Jim Giles, "Internet Encyclopaedias Go Head to Head," *Nature,* December 15, 2005, 438: 900–901.

3. "Encyclopaedia Britannica and Nature: A Response," Nature Online, March 23, 2006, http://www.nature.com/press_releases/Britannica_response.pdf; Andrew Orlowski, "Wikipedia Science 31% More Cronky than Britannica's," *The Register*, December 16, 2005, https://www.theregister.co.uk/2005/12/16/wikipedia_britannica_science_comparison.

4. Eric S. Raymond, *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary* (Beijing: OReilly, 2011).

5. Lev Grossman, "You—Yes, You—Are TIME's Person of the Year," *Time Magazine*, December 25, 2006, http://content.time.com/time/magazine/article/0,9171,1570810,00.html.

6. Caitlin Dewey, "Wikipedia Has a Ton of Money. So Why Is It Begging You to Donate Yours?" *Washington Post*, December 2, 2015, https://www.washingtonpost.com/news/the-intersect/wp/2015/12/02/wikipedia-has-a-ton-of-money-so-why-is-it-begging-you-to-donate-yours/; Andrew Orlowski, "Will Wikipedia Honour Jimbo's Promise to STOP Chugging?" *The Register*, December 16, 2016, https://www.theregister.co.uk/2016/12/16/jimmy_wales_wikipedia_fundraising_promise/.

7. Aaron Mak, "Inside the Brutal, Petty War Over Donald Trump's Wikipedia Page," *Slate Magazine,* May 28, 2019, https://slate.com/technology/2019/05/donald-trump-wikipedia-page.html.

8. Matt Reynolds, "A Bitter Turf War Is Raging on the Brexit Wikipedia Page," *Wired UK*, April 29, 2019, https://www.wired.co.uk/article/brexit-wikipedia-page-battles.

9. Aria Thaker, "Indian Election Battles Are Being Fought on Wikipedia, Too," *Quartz India* May 16, 2019, https://qz.com/india/1620023/aatish-taseers-wikipedia-page-isnt-the-only-target-of-modi-fans/.

10. Nima Nazeri and Collin Anderson, "Citation Filtered: Iran's Censorship of Wikipedia," University of Pennsylvania Scholarly Commons, November 2013, https://repository.upenn.edu/iranmediaprogram/10/.

11. NetBlocks, "Wikipedia Blocked in Venezuela as Internet Controls Tighten," *NetBlocks,* January 28, 2019, https://netblocks.org/reports/wikipedia-blocked-in-venezuela-as-internet-controls-tighten-XaAwR08M.

12. Amit Singhal, “Introducing the Knowledge Graph: Things, Not Strings,” Google Blog, May 16, 2012, https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.

13. Nicholas Vincent, Isaac Johnson, and Brent Hecht, “Examining Wikipedia With a Broader Lens,” *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, https://doi.org/10.1145/3173574.3174140.

14. Caitlin Dewey, “You probably haven’t even noticed Google’s sketchy quest to control the world’s knowledge,” *The Washington Post*, May 11, 2016, https://www.washingtonpost.com/news/the-intersect/wp/2016/05/11/you-probably-havent-even-noticed-googles-sketchy-quest-to-control-the-worlds-knowledge/.

15. “In the Media,” Wikimedia Signpost, May 28, 2016, https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2016-05-17/In_the_media.

16. Andreas Kolbe, “Op-Ed,” Wikimedia Signpost, August 30, 2019, https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2015-12-02/Op-ed.

17. Heather Ford and Mark Graham, “Provenance, power and place: Linked data and opaque digital geographies,” *Environment and Planning D: Society and Space*, *34*(6) (2016): 957–970, https://doi.org/10.1177/0263775816668857; Heather Ford and Mark Graham, “Semantic cities: Coded geopolitics and the rise of the semantic Web,” in *Code and the City, ed. Rob Kitchin and Sung-Yueh Perng (Routledge, 2015).*

18. Rachel Abrams, “Google Thinks I’m Dead,” *The New York Times*, December 16, 2017, https://www.nytimes.com/2017/12/16/business/google-thinks-im-dead.html.

19. Brian Heater, “Are corporations that use Wikipedia giving back?” *TechCrunch*, March 24, 2018, https://techcrunch.com/2018/03/24/are-corporations-that-use-wikipedia-giving-back/.